I/S: A JOURNAL OF LAW AND POLICY FOR THE INFORMATION SOCIETY

Artificial Intelligence Effecting Human Decisions to Kill: The Challenge of Linking Numerically Quantifiable Goals to IHL Compliance

LIEUTENANT COLONEL ALAN L. SCHULLER*

CONTENTS

I. INTRODUCTION

Future military forces will face defeat at the hands of a near-peer adversary if they do not possess the capacity to act and respond faster than a human being. As such, in the interest of national defense, technologically advanced nations will develop weapon systems that

*U.S. Marine Corps; Stockton Center for the Study of International Law, U.S. Naval War College

rely increasingly on artificial intelligence (AI) to effect human decisions. These weapons will sense their environment, process inputs from this and other data, and respond quicker than their human foes are able. Some of the states that create such systems will adhere to the idea that the laws of war establish substantive and ideological standards that must be adhered to, rather than simply serving as a means to an end. During the development of AI-enhanced weapon systems, these countries will face the challenge of scrutinizing the technology behind how AI effects human decisions. Those charged with reviewing the legality of such systems will encounter stark contrasts in the predictability of aspects of AI systems as compared to legacy systems. As a result, lawyers will need to understand how AI arrives at a "decision" and how an AI "decides" differently than humans. One of the critical differences is in the description and attainment of objective performance standards. This paper explores the challenges presented when establishing such standards for AI and describes some situations where AI systems could achieve success or fall short in hypothetical military applications. It concludes that the ability of lethal AI to comply with International Humanitarian Law (IHL) will depend in part on whether the goals assigned to it are amenable to description as numerically quantifiable objective standards.

II. THE DEPENDENCE OF ARTIFICIAL INTELLIGENCE ON NUMERICALLY QUANTIFIABLE GOALS

A. *Background*[1]

The decision to kill may never be functionally delegated to a computer. More specifically, evaluating the lawfulness of a use of force in the context of armed conflict is by definition a human task. Ensuring compliance with the law is a burden that may not be abrogated by surrendering such authorities and capabilities that humans no longer reasonably control the decision to kill. In the process of fielding a weapon system for use in armed conflict, we must therefore inquire into "how confidently we can establish in advance that a weapon system will kill the intended people or classes of people

[1] This section summarizes relevant points from Alan L. Schuller, *At the Crossroads of Control: The Intersection of Artificial Intelligence in Autonomous Weapon Systems with International Humanitarian Law*, 8 HARV. NAT'L SEC. J. 379 (2017).

and destroy the intended objects or classes of objects."[2] The precise technological manner in which the decision to kill might be functionally delegated, however, is difficult to describe with certainty because hypotheses on the matter are often intertwined with assumptions about how future systems will develop that are shaded by institutional bias. It is nevertheless important to describe general characteristics of AI that potentially invite functional delegation. One such aspect is the capacity of AI systems to achieve certain objective performance standards.

B. *Objective Performance Standards*

There are endless ways to define, describe, and evaluate AI. In the context of examining AI-enhanced weapon systems and whether they can comply with the laws of war, however, we are concerned with the *effects* produced by the weapon. The *processes* by which the system arrives at a given outcome are of less concern. Further, holding AI to the arguably low standards sometimes demonstrated by humans is widely considered insufficient. As such, it should be relatively uncontroversial that an AI-enabled weapon must be "evaluated based upon how well it performs to rational and objective standards."[3] Setting a rational standard means we must describe an ideal standard to which we expect the system to perform. In this context, an ideal standard would fall somewhere in between human performance standards and perfection. An objective goal indicates measurable, fact-based standards as compared to subjective personal opinions or judgment. In the context of an AI-enhanced weapon, however, one must pause to further consider the difference between these modes of analysis.

C. *Objective Versus Subjective AI?*

In the context of reviewing a weapon system, an objective standard describes a performance level established as reasonable in the eyes of the law as evaluated by those charged with determining compliance. This means that when a lawyer evaluates a weapon system in order to determine conformity with IHL, she asks whether

---

[2] *Id.* at 416.

[3] *Id.* at 401.

the system is *per se* illegal, and if not, whether a commander could reasonably employ it in accordance with the law of war.[4] This begs the question, however, as to what objective standard of performance is reasonably sufficient to comply with the law. It is at least conceivable that in the future we may need to design AI that can perform to objective standards described at a minute level of detail where we delineate precisely which individuals will be killed for the sake of achieving military advantage.[5] That time, however, has not yet arrived and is beyond the scope of this paper. We may therefore sidestep the matter and be satisfied that for the time being, AI-enhanced weapons will need to perform to a generalized standard that policymakers and lawyers collectively decide is objectively reasonable. That is because the AI must necessarily attempt to effect human judgment as to an objectively reasonable outcome.

There is no such thing as subjective AI. Computers do not possess instincts, values, judgment, or higher consciousness; and they probably never will. AI systems may impressively mimic such human characteristics, but they only do so as a function of advanced programming. When AI finds a coffee shop that is ideal for you, it is because the algorithm behind its software is optimized effectively to achieve its designated goal. The AI does not care about you or your desires or values despite the fact that it may consider those facts when arriving at a recommendation. The answer it provides is therefore by definition objective because the AI was programmed by humans to achieve an established level of performance. The values a system places on the inputs it processes in arriving at an output are the result of computational statistics. To be sure, the goals established for an AI by its programmers may reflect their subjective value judgments, but these may not be imputed to the machine. An AI does not value the

---

[4] Legal reviews of new weapons are conducted as a function of customary international law or as mandated by Article 36 of Additional Protocol I of the Geneva Conventions. *See* Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts art. 36, *opened for signature* June 8, 1977, 1125 U.N.T.S. 3 (entered into force Dec. 7, 1978), https://treaties.un.org/doc/publication/unts/volume%201125/volume-1125-i-17512-english.pdf ("In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party") [https://perma.cc/ZPK6-YXVU] (hereinafter API).

[5] *See Moral Machine*, MASS. INST. OF TECH., (last visited June 21, 2018) http://moralmachine.mit.edu [perma.cc/3SB8-TBXT].

accomplishment of assigned goals any more than a sewing machine values the quality of the clothes it helps create. AI can achieve subjective human standards, but only because it was programmed to do so. It does not possess the capacity to "know" the difference.

Humans, on the other hand, appear at times incapable of divorcing their subjective judgment from the consideration of rational objective standards. For any student of the law, this becomes painfully obvious during schoolhouse discussions struggling through the standard of the "reasonable person." But most of these debates arguably center around whether a standard is a reasonable one, not whether it is possible in theory to establish standards that most would view as objectively reasonable. So we proceed for the sake of argument, therefore, that objective standards may be established that are viewed by the majority of persons as reasonable.

Consider for example the requirement pursuant to IHL to ensure attacks do not produce disproportionate harm to civilians. Pursuant to the principle of proportionality, expected loss of civilian life and damage to property incident to an attack must not be "excessive in relation to the concrete and direct military advantage anticipated."[6] This standard of course begs the question, "What is excessive?" Most military leaders would agree that the aerial bombing of an orphanage would not be justified simply to eliminate one enemy foot soldier on the roof. But a policy on the other end of the spectrum that forbade any attack where one civilian might die would ensure defeat of that hamstrung military force. Any hypothetical within these two extremes, however, inevitably invites endless debate and ultimately results in humans "agreeing to disagree." This does not mean that a rational objective standard is unattainable. It simply means we cannot agree on where to establish the red line. In the past, there was no imperative to set clear standards because the task of deciding what was "excessive" was left to military commanders and rarely second-guessed. Observers have lamented that machines could never possibly be able to replicate human decision-making in this realm. There are many conundrums to consider in answering this complicated question, but a key one amongst them is whether the decision a machine is tasked with is the type of decision that machines are adept at making on our behalf. An important aspect of the inquiry is whether the goal established for an AI is numerically quantifiable.

---

[6] API, *supra* note 4, at art. 51(5)(b); *see also* JEAN-MARIE HENCKAERTS & LOUISE DOSWALD-BECK, CUSTOMARY INTERNATIONAL HUMANITARIAN LAW, 46 (Cambridge Univ. Press 2005).

D. *How Decisions Are Delegated to Machines*

Generally speaking, the capabilities and limitations of any given weapon system depend on what humans empower it to accomplish. Machines do not decide to do anything in the human sense.[7] They follow their programming. But the ways in which machines effect human decisions are more complicated than ever. Modern AI is not predictable in the way legacy systems were and actions taken by AI might evade reverse engineering after the fact. But these problem sets are quite different than possessing the free will, for example, to redefine goals.

I argue that the law forbids humans from creating systems that have such extensive authority and capabilities that we cannot reasonably predict whether they can be employed in accordance with IHL.[8] Capabilities are comprised by the physical attributes provided to the system, such as a vehicle platform, ability to loiter, sensor suites, and types of weapons. Authorities are determined by the computer programs governing the system, which could range from traditional deterministic systems to an advanced deep neural network empowered with *in situ* learning capability. The inquiry regarding AI decision-making in this context does not hinge on how *much* authority or how *many* capabilities we delegate, but instead exactly what *combinations* are provided.

The authorities granted to an AI-enabled weapon system will incorporate goals for the AI to attain, which are described by rational, objective standards. For at least the foreseeable future, AI will require these standards to be described in a numerically quantifiable manner.[9] This means that the goals established for the AI are amenable to quantitative description. For example, "do not kill more than five civilians" is a quantifiable standard, discounting, of course, the question of whether the system is capable of distinguishing between the status of people on the battlefield. On the other hand, "do the right thing" is inherently not a quantitative standard. While we could suggest quantifiable goals that might approximate our vision of what "the right thing" is, these goals would simply serve as numerical proxies for a standard that is by definition qualitative.

---

[7] *See supra* note 1, at 388 n.40.

[8] *See supra* note 1, at 392.

[9] *See* Interview with Machine Learning Experts, OpenAI, San Francisco, CA (Jan. 26, 2018).

Depending on the goal in question, describing rational objective standards in a numerically quantifiable way might be simple or impossible. The use of proxy standards to achieve qualitative goals could suffice in some situations. Often, reactions to hypothetical AI systems are laden with assumptions about the way we believe AI might develop as well as our value judgments about what direction it should take. The reality is, however, that it will depend on the specific tasks we ask the AI to accomplish. The ability of AI-enabled weapon systems to comply with IHL will depend in part on whether the tasks in question are susceptible to being described as numerically quantifiable objective standards. In the following sections, we will explore hypotheticals that will examine this concept in greater detail.

It is worth noting that, by clearly delineating numerically quantifiable standards, we may limit the usefulness of AI in military applications. Amongst the many advantages that AI potentially provides the military is the ability to make completely unpredictable decisions at speeds no human can match. If AI systems are only employed in situations where they can evaluate easily quantifiable factors, they may be so limited as to negate their military advantage. For this reason, the tempting and simple answer that AI should not be used in difficult-to-quantify environments is unsatisfying.[10]

In the sections that follow, we consider two hypotheticals that evaluate the application of possible future AI-enabled weapon systems in the context of particular IHL rules. The examples are illustrative and not exhaustive and do not consider the full panoply of IHL rules that might apply to the hypothetical systems. They will serve, however, to illuminate some of the challenges faced in translating the military application of an AI-enabled weapon into numerically quantifiable goals.

E. *Numerically Quantifiable Goals to Attain Rational Objective Standards*

The ways modern AI systems accomplish goals seem to attenuate the link between human decision and machine action. Sometimes even AI programmers do not understand, for example, why neural

[10] It is also worth mentioning that all of the hypotheticals in this paper relate to lethal weapon system. Without question, AI presents opportunities to ensure our national security in non-lethal contexts such as intelligence analysis and information operations, but these non-lethal contexts are beyond the scope of this paper.

networks operate in the manner they do or how such systems might accomplish an assigned goal. Further complicating the matter, although humans may define goals for AI, there are certain kinds of goals that AI are particularly adept at achieving and others at which they perform quite poorly. Taken together, these points invite us to explore the kinds of military applications in which AI might fulfill or fall short of legal standards.

As a point of departure, consider the difference in the search algorithms powering the websites of Google and YouTube. Both sites have search engines that are managed by subsidiaries of Alphabet Corporation. But the algorithms behind the two sites are optimized to achieve very different results. In short, the goal of Google is to direct users to the website that most accurately conforms to what you are searching for, while the goal of YouTube is to get users to click on as many videos as possible.[11] As a result, users may be presented with drastically different suggestions depending on which site they visit. That being said, both websites provide opportunities for their designers to numerically quantify success. For Google, if the user clicked on the first result from their search (setting aside entirely the issue of advertising) then the AI performed successfully. If the user was forced to scroll to later results, the performance was suboptimal. For YouTube, success simply means continuous clicks. The more clicks, the better the results. But not all goals are as simple to express in numerical terms.

Suppose YouTube was tasked by its leadership to respond to search queries with what a user "should" see in response to a given input. The computer programmers at YouTube would be faced with the task of overhauling their search algorithm in order to return videos that reflected the value judgments made by that company as to an appropriate response. The most obvious hurdle would be to establish what viewers "should" see. For example, if a user types "white power forever," is it YouTube's role simply to provide the obvious but societally discouraging output of extremist videos? Or should the website steer users to films intended to eradicate xenophobia? This is of course the most thorny and inflammatory aspect of the question, which has implications far beyond subjective value judgments (and hence the scope of this paper). But the second

---

[11] Jack Nicas, *How YouTube Drives People to the Internet's Darkest Corners*, WALL ST. J. (Feb. 7, 2018), https://www.wsj.com/articles/how-youtube-drives-viewers-to-the-internets-darkest-corners-1518020478 (contrasting results from the two search engines) [https://perma.cc/W929-KHVW].

difficulty encountered when tailoring the AI to achieve this goal is in defining success. In other words, how would YouTube know that it had accomplished its goal of directing users to the videos that they "should" see? While the problem may not be intractable, it is without question more complicated than simply "more clicks is better." We could devise standards aimed at approximating success in achieving our subjective goals, such as downward trends in the number of searches for "white power forever" and related queries. But such standards seem like an unsatisfying means to quantify success. Less searches for such unsavory topics does not necessarily mean the underlying societal issues are resolved. There are simply some goals that are difficult to quantify in numerical terms. These are the goals at which AI-enabled systems will be less adept at achieving to satisfaction.

III. MILITARY APPLICATIONS OF LETHAL ARTIFICIAL INTELLIGENCE AND THE CHALLENGE OF SUBJECTIVE IHL STANDARDS

In this section we explore two hypothetical military applications of lethal AI. We begin by examining a system that leverages the strengths of AI's ability to target objects based on numerically quantifiable standards. The second example delineates another end of the spectrum, where AI is likely unable to comply with IHL due to the subjective nature of the goals assigned.

IHL is the *lex specialis* that governs parties to an armed conflict. Its core principles are military necessity, distinction, proportionality, and preventing unnecessary suffering. Military necessity means that a belligerent may attack targets that are indispensable for defeating the enemy so long as they are not otherwise illegal.[12] The principle of distinction holds that only military targets may be attacked, and commanders must proactively determine if potential targets are civilians or combatants, and then attack only combatants.[13] Proportionality is a concept that balances unintentional harm to

[12] API, *supra* note 4, at art. 51(5)(b) (attacks on targets that would produce a "concrete and direct military advantage," and are not otherwise unlawful, are not prohibited); API, *supra* note 4, at art. 52(2) (targets are persons and objects "which by their nature, location, purpose, or use make an effective contribution to military action" and whose destruction or neutralization "offers a definite military advantage.").

[13] *Id.* at art. 52(2) ("Attacks shall be limited strictly to military objectives."); HENCKAERTS & DOSWALD-BECK, *supra* note 6, at R.7.

civilians and their property with the military advantage of attacking a target. The collateral damage from an attack cannot be clearly excessive in relation to the military advantage anticipated from the attack.[14] The principle of unnecessary suffering prohibits the use of weapons that by their nature cause unnecessary suffering and also the use of lawful weapons in a manner that is intended to cause unnecessary suffering.[15] While these principles will not be comprehensively evaluated in the following hypotheticals, it is important to not overlook the fact that each may present unique challenges when applied in the context of lethal AI.

A. *Example One: Quantifying Distinction*

An unmanned submarine powered by AI could search for and destroy enemy submarines during an international armed conflict. Such a system would leverage technology that is already in existence or could be produced in the reasonably foreseeable future that employs goals, which are numerically quantifiable. Importantly, the nature of the operating environment makes applicable legal standards simpler to achieve.

A submarine-hunting platform would rely on multiple technologies that take advantage of the strength of AI. It would leverage advanced forms of the non-visual object recognition and classification systems in use today. The system could process intercepts in the form of signals and acoustics similar to the manner in which current submarines do. It could also process friendly identifiers in order to avoid fratricide as is commonplace among modern militaries, but an AI could process all of this information in order to arrive at a targeting decision far faster than a human. As additional sensor systems are developed that surpass today's technology, the AI will arguably be better suited to incorporate voluminous amounts of data into action. These types of data can be translated into statistical expressions of certainty. For example, given X signals intercepts, Y acoustic signature, Z undersea maneuvers, ruling out other identifiers to include sea life and friendly forces, the AI produces a result that it is a percentage (of 100 for instance) certain

---

[14]API, *supra* note 4, at art. 51(5)(b); HENCKAERTS & DOSWALD-BECK, *supra* note 6, at R.14.

[15]API, *supra* note 4, at art. 35(2)–(3); Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 1996 I.C.J. 226, ¶ 78 (July 8); HENCKAERTS & DOSWALD-BECK, *supra* note 6, at R.70.

that the object it detected is an enemy submarine. If the statistical certainty is above a cutoff value established by humans, the AI will attack. It could also have other hard rules in place to stop the attack, for example if the enemy submarine was in a harbor or in close proximity to an unidentified vessel.

The counter-argument to this type of system is that even if you accept all of this as true, it will remain imperative to have a human onboard or in communication in order to at least override mistakes by the AI. This human interaction comes at a significant cost, however, to the overall effectiveness of the system.

The presence of humans aboard a submarine significantly degrades the optimal performance of that platform. Unmanned submarines operating with AI would likely be able to outmaneuver and destroy manned submarines during future armed conflict at sea. First, humans are noisy. A machine that does not need to communicate is by definition quieter. Second, humans need to breathe and are easily crushed. This means manned submarines simply cannot dive as deep as unmanned versions. Depth often means survivability and tactical advantage. Third, humans eventually need to surface to resupply. An AI does not, at least not at the same intervals. Fourth, driving submarines is difficult and if you do it wrong lots of sailors die. An AI is simply a machine whose loss is insignificant beyond its military utility and dollar value, and this fact encourages investment in such systems and possibly reduces cost. Of course, none of this means that modern militaries should or will replace all manned submarines with unmanned systems, it simply identifies that there is an arguably significant military advantage to be gained by the development of AI-enabled unmanned submersibles.

In this hypothetical, the ability of the system to comply with IHL and the Law of Naval Warfare is simplified due to the nature of the operating environment. There simply are not that many civilians or civilian objects operating at the depth of military submarines. Consider the IHL principle of distinction. In armed conflict, this limits belligerents to attacking targets that are valid military objectives.[16] As a subset of distinction, the imperative to take precautions in the attack requires that a belligerent take active steps to determine whether persons are civilians or combatants and to direct operations only

[16] API, *supra* note 4, at art. 52(2) ("Attacks shall be limited strictly to military objectives."); HENCKAERTS & DOSWALD-BECK, *supra* note 6, at R.7.

against combatants.[17] The task of distinguishing combatants from civilians in the land domain is challenging. This is especially so in cluttered areas such as cities or during conflicts in which the enemy intentionally commingles with civilians, which has been the recent experience of most of the United States military for well over a decade of conflict. But the military who prepares for the last war will lose the next one. Distinction in the undersea context is simply not as challenging as on land due to the dearth of human presence. As such, the likelihood that a miscalculation will kill civilians is correspondingly reduced. Further, AI systems in the undersea context can be provided the technological capability to distinguish in ways that land systems are not currently able. Using signals and acoustics as well as other data, an AI could come to a conclusion that was statistically reasonable and based on numerically quantifiable data points that an undersea contact was an enemy submarine. These types of quantifiable factors are more difficult to discern in a cluttered urban land context, for example in positively identifying an enemy fighter and distinguishing that person from a nearby civilian. As such, AI arguably presents opportunities in this realm due to its ability to distinguish using numerically quantifiable and rational objective standards. Such a system might not ever leave port during actions short of international armed conflict, but it could prove decisive in the event of naval warfare with a peer competitor.

B. *Example Two: Quantifying Unnecessary Suffering*

The IHL prohibition against causing unnecessary suffering provides contrast to the previous application of the distinction principle. At least in the example above, an AI is faced with making a factual determination that will classify contacts as legitimate targets or not. Either an underwater object is an enemy combatant's submarine or it is not. In the more complicated context of urban land warfare, it is the factual determination of civilian or combatant that is currently more difficult for the AI to discern. But in either context, the distinction challenge is a binary one that is complicated only by the clutter of the operating environment and the ambiguity of human behavior. In the undersea context of the submarine hunter, an AI might reasonably be able to apply the distinction principle because it can detect and analyze objective numerically quantifiable information.

---

[17]API, *supra* note 4, at art. 57; HENCKAERTS & DOSWALD-BECK, *supra* note 6, at R.15–21.

In the land context the same principle might theoretically apply but current technology appears unsuited to make such fine grained distinction on its own. The challenge of AI applying the concept of unnecessary suffering is different *in kind*, however, because the principle is less amenable to numerically quantifiable rational goals.

The prohibition against causing unnecessary suffering makes unlawful the use in armed conflict of weapons that by their nature cause unnecessary suffering and the use of lawful weapons in a manner that is intended to cause unnecessary suffering.[18] Importantly, there is no simple objective test to determine whether the use of a weapon would constitute unnecessary suffering.[19] The first aspect of the principle is addressed during legal review of proposed weapons and is contextually specific to the system being evaluated. The contours of specific prohibitions established by customary law and treaty are beyond the scope of this article. As such, this discussion will generally focus on the second prong regarding employment of weapons already deemed not *per se* unlawful.

In theory, the goal of preventing unnecessary suffering is a noble one. Most would agree in principle that no more suffering should be caused to combatants than is necessary to obtain military victory. In application, however, the principle is riddled with subjective judgment. Consider two simple examples. First, if a soldier bayonets an enemy, does she cause unnecessary suffering if she twists and turns the bayonet after stabbing the enemy? Some might argue that the additional pain caused by twisting the weapon is simply unnecessary as the victim has already been wounded and is most likely out of the fight. On the other hand, the soldier is permitted under the laws of war to continue attacking until the enemy is dead, assuming the enemy does not surrender or is not clearly *hors de combat*. The action of twisting the blade will help ensure a fatal wound to the enemy, and, as such, she is arguably well within the law to twist and turn and stab again until the enemy is dead. By way of a second example, consider the use of an artillery barrage of high explosive rounds mixed with white phosphorus shells, commonly referred to as an HE/WP fire mission. Some would argue that the combination of these rounds creates unnecessary suffering because the white phosphorous causes horribly painful burns on enemy soldiers that are exposed to the

---

[18] API, *supra* note 4, at art. 35(2)–(3); Advisory Opinion, *supra* note 15, at ¶ 78; HENCKAERTS & DOSWALD-BECK, *supra* note 6, at R.70.

[19] GARY D. SOLIS, THE LAW OF ARMED CONFLICT 271–72 (Cambridge Univ. Press 2010).

barrage, and that the soldiers could be killed more humanely by using only HE rounds or other less painful weapons. On the other hand, the employment of HE/WP is highly effective at destroying enemy fuel depots because after the HE rounds pierce fuel containers, the WP rounds light the exposed fuel on fire and destroys the target far more efficiently than one type of round alone. The suffering caused to attendant enemy soldiers is arguably an unfortunate byproduct of a completely lawful attack. In sum, when one of the core functions of the military is to kill human beings during armed conflict, any discussion about what suffering is unnecessary is fraught with subjectivity.

These are merely two very simple examples that have plausible arguments on both sides. We might spend hours debating them without arriving at consensus regarding whether they violate the principle of preventing unnecessary suffering. The standard is inherently subjective and in application nearly impossible to divorce from one's biases, be they cultural, institutional, or otherwise. As such, application of this principle may prove highly problematic for an AI because it is less amenable to description in numerically quantifiable terms. In other words, it is quite difficult to approximate using rational objective goals that are defined by numerical standards.

Since machines are not capable of forming intent, a more detailed inquiry into the matter starts with how AI systems could be deployed by humans in a manner intended to cause unnecessary suffering. The evaluation would hinge on parsing out that suffering which is a byproduct of defeating the enemy from that suffering which is excessive, and thus, unnecessary to secure victory. One could of course conjure up AI systems that would violate this principle. Suppose that a military developed a "Pain Bot" that was designed to learn how to kill enemy soldiers as slowly as possible without allowing itself to be captured or destroyed. Of course, this would be *per se* illegal, and it would not be created by any country that was dedicated to the rule of law and IHL principles. Such a sophomoric example aside, the analysis becomes complex.

Suppose instead that a country developed a robot for deployment in urban combat called "Surrender Bot." It is equipped with a high-power laser that can cut through an enemy soldier's body armor. The country intends to deploy the robot in close quarters as the first system to enter enemy held buildings. One of their objectives in doing so is to kill as few enemy soldiers as is necessary in order to achieve the most efficient military victory possible, thus encouraging post-conflict reconciliation. Most notably, the robot is equipped with AI

that allows it to learn what employment of its laser is most effective at obtaining the surrender of the enemy. The system is trained in a laboratory on this goal, but continues to learn *in situ* on the battlefield. Its operators do not know ahead of time which enemy soldiers will be targeted or where on the enemy soldiers the AI will direct its laser. After deployment, the Surrender Bot proves highly effective at compelling enemy forces to surrender. Sometimes it chops off the enemy's legs and other times it kills the enemy instantly by shooting them in the forehead. If it detects an enemy leader it generally kills that person first. In some instances, Surrender Bot will deliberately injure, but not kill an enemy in order to cause that fighter to scream and writhe in pain, thus encouraging the surrender of the enemy fighters in close proximity, but due to hard programming rules the AI never attacks any soldier that has surrendered or is wounded.

We sidestep the issue of whether Surrender Bot is unlawful *per se* because by its nature it causes unnecessary suffering without conceding the matter. That *ex ante* determination does not implicate the ability of the AI to achieve numerically quantifiable goals in the operating environment. The second prong of the principle, however, requires those designing the AI to describe goals that can be quantified in a way that makes them amenable to application by AI. In other words, how can we provide the AI with objective standards that will enable it to determine when the suffering it creates is unnecessary under the circumstances? Given the subjective nature of this analysis as delineated above, this task appears a significant challenge.

Surrender Bot's designers would be perplexed by multiple and competing goals for the system. It would need to be empowered to kill the enemy.[20] We assume simply for the sake of argument, again without conceding the matter, that the AI could detect and identify enemy fighters. So this begs the question, "which soldiers should be killed?" The AI could be programmed with great discretion in this matter (i.e. there is no limit to the number of enemy soldiers that may be killed until victory is achieved) or it could be bounded significantly in countless ways. For example, it could have a hard rule that the enemy must be wounded first and allowed the opportunity to surrender, and only killed if they refused to surrender. The next question is whether the AI should be limited in the other manners in which it can use its laser. Is it allowed to shoot the enemy in the

[20] Other options could certainly include non-lethal means of incapacitating the enemy but are thus beyond the scope of this article.

kneecaps? In the genitals? As a preliminary matter, it would not be allowed to use the laser to blind the enemy soldiers.[21] Beyond that arguably arbitrary limitation, programmers would need to work with lawyers to delineate what kinds of suffering it might create that were unnecessary to achieve victory. This process would be a nightmare because if the machine is not specifically forbidden by hard rules from taking a particular action, we must assume that the AI may use that option without any regard for what a human might do under the circumstances. Machines do not have common sense, values, morality, decency, sympathy, empathy, or any of the other traits that make some acts by humans generally less likely. We cannot tell an AI, "do the right thing," or "you will know it when you see it." So in programming the AI, we must attempt to approximate in quantifiable terms what we mean by "unnecessary suffering." Given the subjective nature of this principle, that task might prove highly problematic.

One solution might be to dramatically limit actions the AI could take. This simple solution appears elegant at first. The system could be designed to either kill instantaneously or wait. It would not be allowed to take any other actions, thus sidestepping the question of whether its behavior creates unnecessary suffering. But this solution fails entirely to leverage the strength of AI. Learning systems can adapt and make decisions faster than humans and behave in ways that would be unpredictable to the enemy. If their capacity is bound so strictly that it becomes highly predictable, the machine becomes deterministic and not significantly more useful than legacy systems. It would also arguably be less likely to achieve the goal of killing the fewest soldiers possible while securing victory. If it defaulted instead to no action, it would be less survivable on the battlefield. This kind of system might still be useful in achieving military victory, but it would not further the goal of preventing unnecessary suffering.

On the other hand, if the AI was provided broad discretion, we must anticipate that it will behave in a way that no human would.[22] Perhaps the AI might learn to target the weapons in the hands of

---

[21] *See* Protocol on Blinding Laser Weapons, *opened for signature* Oct. 13, 1995, 1380 U.N.T.S. 370 (entered into force July 30, 1998), https://treaties.un.org/doc/Treaties/1995/10/19951013%2001-30%20AM/Ch_XXVI_02_ap.pdf [https://perma.cc/6AXE-CW9M].

[22] *See* Jack Clark & Dario Amodei, *Faulty Reward Functions in the Wild*, OPENAI (Dec. 21, 2016), https://blog.openai.com/faulty-reward-functions/ ("Reinforcement learning algorithms can break in surprising, counterintuitive ways.") [http://perma.cc/4T3L-P8J3].

enemy fighters. This would be highly desirable, as the AI would essentially be able to force the enemy to surrender without causing any suffering. Alternatively, the AI might learn to surgically slice off small portions on multiple parts of the enemy soldiers' body, causing extreme pain and incapacitating enemy soldiers and forcing their surrender. Or it might lobotomize them all. We simply could not assume that any option was off the table for the system unless we programmed it to be off the table.

Even more to the true dilemma, equipping the system to make determinations on its own as to what suffering was unnecessary in the absence of hard rules seems intractable for the foreseeable future. The challenge of approximating subjective standards for unnecessary suffering through hard programming boundaries pales in comparison with attempting to equip AI with the capability of establishing suffering as unnecessary on its own. As described above, machines do not possess any of the human characteristics that underpin our desires to limit the suffering of others. As such, if the AI was programmed to determine what suffering was unnecessary, it would need numerically quantifiable goals that quantified suffering and established limits in relation to military objectives. This would arguably be more complicated to quantify than decisions as to the proportionality of attacks[23] because there is no clearly established benchmark against which to weigh the suffering of the enemy. Suffering is omnipresent in armed conflict. Every military attack causes considerable suffering, the contours of which defy simple definition. Minor suffering could be unnecessary while massive suffering could be well justified under IHL. Under the circumstances described, it appears that AI will be poorly suited to the task of evaluating suffering on its own for the foreseeable future. As such, those designing AI-enabled weapon systems will need to be aware of this limitation and account for it during the design of the system.

IV. CONCLUSION

AI is a highly technical field that has seen dramatic advances as well as wildly overoptimistic visions for its future. It is a safe assumption that future military forces will need to leverage the power

[23] Proportionality under IHL means that the anticipated loss of civilian life and damage to property incidental to attacks must not be "excessive in relation to the concrete and direct military advantage anticipated." API, *supra* note 4, at art. 51(5)(b); HENCKAERTS & DOSWALD-BECK, supra note 6, at R.14.

of AI in order to dominate the battlespace, whatever form that technology takes. While the law may not need to adapt to AI, lawyers applying the law must understand the technology involved more deeply than perhaps was necessary in the past. For the lawyer who applies the law without respect to the facts provides no service to his client.

Further, we must evaluate every weapon system under development and during legal review on its particular facts and merits. Sweeping statements about whether AI-enabled weapon systems will be able to comply with IHL do little to further informed discourse on the subject. Some systems we could design today with advanced AI incorporated in significant ways will be perfectly lawful. Other concepts are simply incongruent with the strengths and limitations of AI. As lawyers continue to develop expertise in AI, we will be better equipped to facilitate processes that will make these fine grain distinctions. This expertise will be integral to ensuring national security as well as adherence to the rule of law and humanitarian ideals.